Anticipating Diabetes: Machine Learning in Healthcare.

Introduction: -

This report aims to examine a given dataset and create a diabetes-predicting model. Diabetes prediction is critical in healthcare because it allows for the early identification of persons at risk and allows for appropriate intervention and control. We want to identify patterns and linkages within the information that will enable us to properly forecast the existence of diabetes by using the capabilities of data analysis and machine learning techniques. This study has the potential to enhance patient outcomes and the efficient use of healthcare resources.

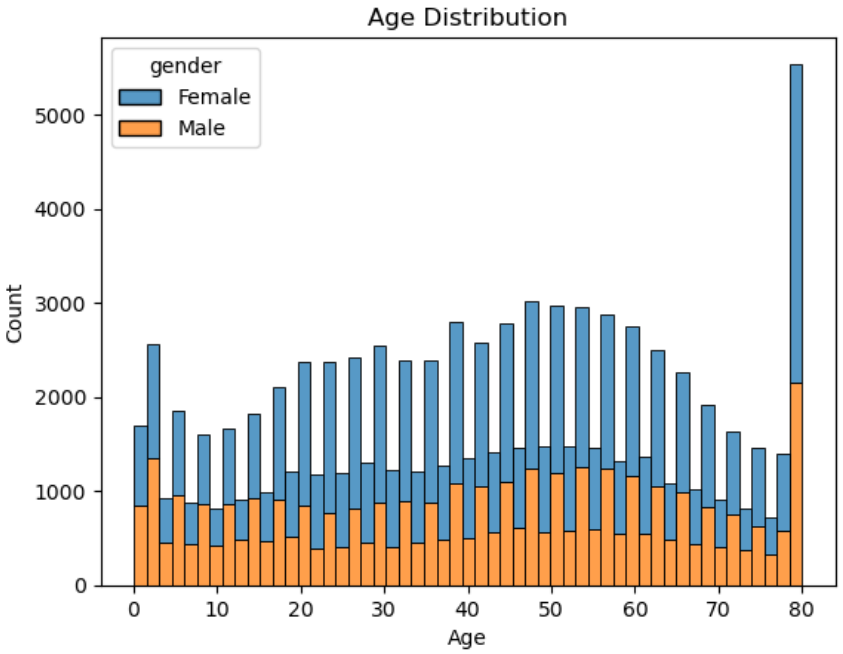
Data Import and Initial Analysis: -

To begin our study, we imported the dataset into Python using the pandas module. This library assists us in exploring and comprehending the data. We also used packages such as numpy, matplotlib, and seaborn to produce visualisations. The dataset is made up of 100,000 rows and 9 columns. Text (for gender and smoking history), numbers with decimals (for age, BMI, and HbA1c level), and full numbers (for hypertension, heart disease, blood glucose level, and diabetes) are among the data types in the columns. We made sure to delete any duplicate entries, so our dataset now has 96,128 distinct rows.

Exploratory Data Analysis: -

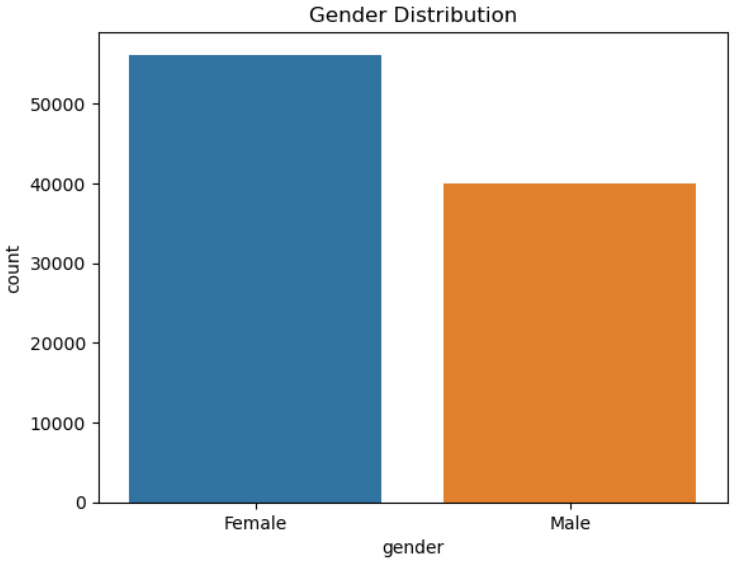
Histogram for the age column: -

The Seaborn library was used to create a histogram to analyze the age distribution in our dataset. The histogram displays the number of people in different age groups and includes a hue parameter to identify gender disparities. The x-axis represents age, while the y-axis shows the number of people. This histogram allows us to identify patterns or trends in the age distribution of our sample.



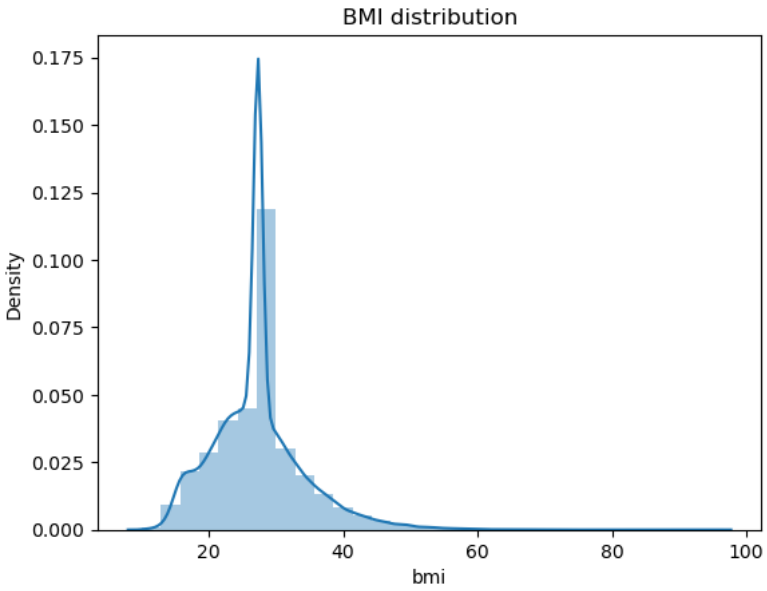
Histogram for the gender column: -

The seaborn library was used to create a count plot to analyze the gender distribution in our dataset. The plot displays the number of people in each gender category, allowing us to identify any gender imbalances or patterns. The x-axis represents gender groups, while the y-axis represents the number of people. This visualization provides valuable insights into the gender makeup of our sample.



Distribution for BMI Column: -

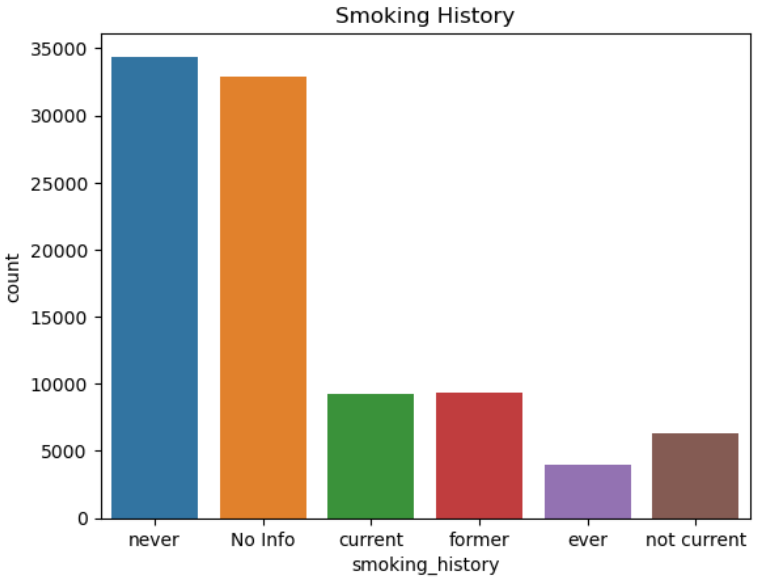
The "BMI Distribution" picture is a histogram that depicts the distribution of BMI values in a dataset. It is made with the distplot function from the seaborn library, with the x-axis indicating the range of BMI values and the y-axis reflecting the frequency of individuals falling within each range. The histogram is separated into bins, and each bar represents a different range of BMI values.



The term "BMI Distribution" expresses the distribution of BMI values within the dataset efficiently, helping visitors to grasp the prevalence of various BMI ranges among the population.

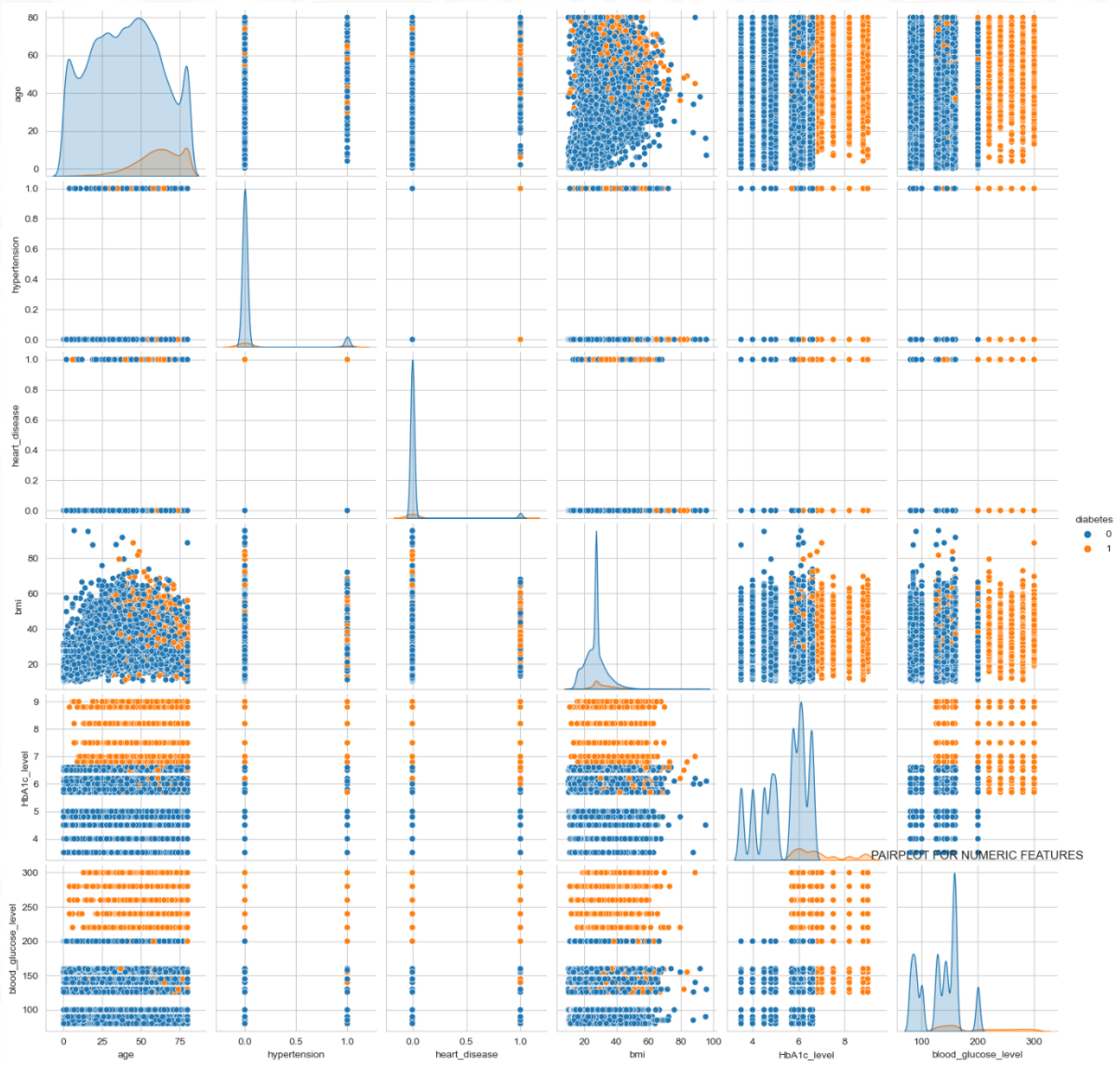
Countplot to Smoking History: -

The image presented, labeled "Smoking History," is a bar chart that depicts the distribution of smoking history in a dataset. The graphic is made with the countplot function of the seaborn library and the smoking history values from the dataset as input. The x-axis shows the various categories of smoking history, while the y-axis shows the number of people who fall into each



group. Each bar in the chart represents a different smoking history category, and its height represents the number of people who fall into that group. The title "Smoking History" clearly describes the substance of the picture. This graphic successfully communicates the distribution of smoking history within the dataset, helping readers to comprehend the prevalence of various smoking history groups among the patients.

Pairplot for Numeric Features: -



The picture "Pairplot for Numeric Features" is a grid of scatter plots that depicts the correlations between numeric features in a dataset. The data points are color-coded depending on the 'diabetes' column

and created using the seaborn library's pairplot function, allowing for discrimination between persons with and without diabetes. One characteristic is represented by the x-axis, while another is represented by the y-axis. The graphic depicts patterns and relationships within the dataset efficiently, allowing users to spot probable correlations or trends.

Data Preparation: -

During the data preparation phase, the smoking history column was simplified by replacing comparable values with new categories such as "non-smoker," "past-smoker," and "current." The gender and smoking history columns were encoded once, transforming categorical data into binary indications for machine learning techniques. The dataset was partitioned into features (X) and target variables (y), allowing the model to learn patterns and predict outcomes depending on the target variable. This distinction aids in the analysis of correlations between characteristics and target variables, resulting in more accurate predictions.

Correlation Analysis: -

To analyze the relationships between variables and gain insights into their strength and direction of associations, a correlation matrix was computed. This matrix was then visualized using a heatmap, which provides a quick and intuitive understanding of the relationships between variables. Additionally, the correlation of each feature with the target variable (diabetes) was analyzed and visualized. This analysis highlights the variables that have the strongest influence on predicting diabetes.

Model Building and Evaluation: -

The Random Forest Classifier method is a common choice for dealing with complicated interactions and unbalanced datasets, which makes it appropriate for diabetes prediction. Optimizing the Random Forest Classifier hyperparameters using GridSearchCV is an excellent strategy since it exhaustively searches for the optimum combination of hyperparameters based on a defined score criteria.

The Synthetic Minority Over-sampling Technique (SMOTE) was used to solve the issue of class imbalance. To balance the dataset, SMOTE creates synthetic samples for the minority class (diabetes), which can improve the model's performance.

Various measures, including as accuracy, precision, recall, and F1-score, were employed to assess the model's performance. These metrics give a full assessment of the model's predictive ability, taking both correct and erroneous predictions into consideration.

A confusion matrix was also displayed to show the model's performance. The confusion matrix displays the number of true positives, true negatives, false positives, and false negatives, providing information about the model's ability to categorize cases properly.

Overall, developing and assessing a diabetes prediction model requires employing the Random Forest Classifier, tuning hyperparameters, applying SMOTE, and evaluating the model's performance using multiple metrics and a confusion matrix.

Conclusion: -

The study reported in this report indicates the Random Forest Classifier's efficiency in predicting diabetes based on the given dataset.

The model attained excellent accuracy and demonstrated good precision, recall, and F1-score, demonstrating its potential for diagnosing diabetic people.

This study's findings have importance in healthcare since early identification and intervention can improve patient outcomes and lessen the burden of diabetes-related comorbidities.

LinkedIn : - Sidharth. A

GitHub: - Sidharth. A